Analysis of the finite precision s-step biconjugate gradient method

Erin Carson
James Demmel

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2014-18
http://www.eecs.berkeley.edu/Pubs/TechRpts/2014/EECS-2014-18.html

March 13, 2014

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
13 MAR 2014		**00-00-2014 to 00-00-2014**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Analysis of the finite precision s-step biconjugate gradient method	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of California at Berkeley,Electrical Engineering and Computer Sciences,Berkeley,CA,94720**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

We analyze the s-step biconjugate gradient algorithm in nite precision arithmetic and derive a bound for the residual norm in terms of a minimum polynomial of a perturbed matrix multiplied by an ampli cation factor. Our bound enables comparison of s-step and classical biconjugate gradient in terms of ampli cation factors. Our results show that for s-step biconjugate gradient the ampli cation factor depends heavily on the quality of s-step polynomial bases generated in each outer loop.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**18**	

Acknowledgement

Research partially funded by DARPA Award Number HR0011-12-2-0016, the Center for Future Architecture Research, a member of STARnet, a Semiconductor Research Corporation program sponsored by MARCO and DARPA, and ASPIRE Lab industrial sponsors and affiliates Intel, Google, Nokia, NVIDIA, Oracle, and Samsung. Any opinions, findings, conclusions, or recommendations in this paper are solely those of the authors and does not necessarily reflect the position or the policy of the sponsors.

ANALYSIS OF THE FINITE PRECISION S-STEP BICONJUGATE GRADIENT METHOD

ERIN CARSON AND JAMES DEMMEL

Abstract. We analyze the s-step biconjugate gradient algorithm in finite precision arithmetic and derive a bound for the residual norm in terms of a minimum polynomial of a perturbed matrix multiplied by an amplification factor. Our bound enables comparison of s-step and classical biconjugate gradient in terms of amplification factors. Our results show that for s-step biconjugate gradient, the amplification factor depends heavily on the quality of s-step polynomial bases generated in each outer loop.

1. Introduction. Krylov subspace methods (KSMs) are a class of iterative algorithms commonly used to solve the linear system $Ax = b$. In classical KSM implementations, in iteration n, the updates to the solution x_{n+1} and residual r_{n+1} consist of one or more sparse matrix-vector multiplications (SpMVs) and vector operations in each iteration. On modern computer architectures, the performance of these operations is *communication-bound*; the movement of data, rather than the computation, is the limiting factor.

Communication-avoiding KSMs (CA-KSMs), based on s-step formulations ([4, 6, 8, 12, 27, 23, 24]), reduce the total communication cost by a factor of $O(s)$ by performing $O(s)$ computation steps per communication step (see, e.g., [3, 7, 13]). This asymptotic reduction in communication cost yields significant speedups in practice for many problems [19].

Although CA-KSMs are mathematically equivalent to their classical counterparts, their finite precision behavior may differ. It has been empirically observed that the rate of convergence of CA-KSMs deviates further from the convergence of the classical method as s increases, and that the severity of this deviation is heavily influenced by the polynomials used for the s-step Krylov bases (see, e.g., [3, 13, 14, 1]).

In this work, we derive Lanczos-type matrix recurrences governing the s-step biconjugate gradient method (BICG) in finite precision arithmetic, which demonstrates the algorithm's relationship to classical BICG. Using the recurrence, we extend the results of Tong and Ye for classical BICG [25] to derive an upper bound on the norm of the updated residual in finite precision s-step BICG. Our bound provides an analytical explanation for commonly-observed convergence behavior of s-step BICG.

2. Related work. We briefly outline the available literature on relevant topics, namely the analysis of KSMs in finite precision and s-step KSMs.

2.1. Analysis of finite precision Krylov methods. There are two primary effects of roundoff error in finite precision KSMs: the maximum attainable accuracy of the solution is decreased, and convergence may deteriorate. Much research has been devoted to better understanding this behavior, and to devise more robust and stable algorithms.

An upper bound on the maximum attainable accuracy for finite precision KSMs, limited by the deviation of the Lanczos residual from the true residual, was obtained by Greenbaum [10]. Greenbaum proved that this bound can be given a priori for methods like CG, but cannot be predetermined for methods like BICG, which can have arbitrarily large intermediate iterates. There are also techniques for alleviating this loss of accuracy, namely, residual replacement strategies, where the computed residual is replaced by the finite precision evaluation of the true residual at carefully

chosen iterations (see, e.g., [22, 26]. In this way, agreement between the true and computed residual is maintained to within a factor of $O(\epsilon)$.

In [11], Greenbaum proved backward stability of the finite precision CG algorithm, by showing that the computed Ritz values lie in small intervals around the eigenvalues of A. There are many other analyses of the behavior of various KSMs in finite precision arithmetic (see, e.g. [17, 16, 25]). The reader is also directed to the bibliography in [20].

Our analysis is most closely related to the work of Tong and Ye [25]. The authors derived a bound for the residual norm of classical BICG in finite precision, expressed as the product of a minimum polynomial of a perturbed matrix and an amplification factor. Our analysis generalizes the work of Tong and Ye to the s-step BICG case.

2.2. s-step Krylov subspace methods. The first instance of an s-step method in the literature is Van Rosendale's s-step CG [27]. Van Rosendale's implementation was motivated by exposing more parallelism using the PRAM model. Chronopoulos and Gear later created the s-step GMRES method with a similar goal [5]. Walker used s-step bases to improve stability in GMRES by replacing the modified Gram-Schmidt orthogonalization process with Householder QR [28]. These authors used monomial bases, and found that convergence often could not be guaranteed for $s > 5$. It was later discovered that this behavior was due to the inherent instability of the monomial basis, which motivated research into the use of other bases for the Krylov subspace.

Hindmarsh and Walker tried a scaled monomial basis to improve convergence [12], but saw only minimal improvement. Joubert and Carey implemented a scaled and shifted Chebyshev basis which led to more accurate results [14]. Bai et al. improved convergence using a Newton basis [1]. Although successively scaling the basis vectors can lower the condition number of the basis, this computation reintroduces communication dependencies. Hoemmen solved this using a novel matrix equilibration and balancing approach as a preprocessing step, which often alleviated the need for scaled basis vectors [13].

Hoemmen et al. [7, 13, 19] derived communication-avoiding variants of Lanczos, Arnoldi, CG and GMRES. We use *communication-avoiding* to specifically refer to s-step variants implemented using the communication-avoiding matrix powers kernel, which applies to well-partitioned sparse matrices (see, e.g., [18]). Derivations of communication-avoiding variants of nonsymmetric Lanczos-based KSMs, such as BICG, CGS, and BICGSTAB can be found in [3].

2.3. s-step BICG. We briefly review s-step BICG for solving $Ax = b$, where $A \in \mathbb{R}^{N \times N}$ (see Alg.1). Note that this overview is meant for the familiar reader; in the interest of space, we defer to numerous other works on the topic, such as [3, 4, 5, 7, 13, 15, 27, 24]. For simplicity, we assume A is full rank.

Throughout the remainder of the paper, $0_{i,\ell}$ denotes a zero matrix of size $i \times \ell$ and 0_i is a column vector of i zeros. We use I to denote the square identity matrix; dimensions are either given as a single subscript, or are implicit from context. We use e_i to denote the i^{th} column of appropriately sized I.

In each outer loop k of s-step BICG, we generate Krylov bases with the current search direction and residual vectors, p_{sk} and r_{sk}, which we denote as $\underline{V}_k^p$, having basis length $s+1$, and $\underline{V}_k^r$, having basis length s. The basis vectors, or columns of $\underline{V}_k^p = \left[v_{k,0}^p, \ldots, v_{k,s}^p\right]$, are generated by the three-term polynomial recurrence

$$v_{k,i+1}^p = \gamma_i \left(A - \theta_i I\right) v_{k,i}^p + \sigma_i v_{k,i-1}^p \tag{2.1}$$

Algorithm 1 s-step BICG

x_0, $r_0 = \tilde{r}_0 = b - Ax_0$, $p_0 = \tilde{p}_0 = r_0$, $k = 0$
while not converged **do**
  Calculate $\underline{V}_k^p$, $\underline{V}_k^r$, $\underline{V}_k^{\tilde{p}}$, $\underline{V}_k^{\tilde{r}}$, $\underline{B}_k$
  $\underline{V}_k = [\underline{V}_k^p, \underline{V}_k^r]$, $\tilde{\underline{V}}_k = [\underline{V}_k^{\tilde{p}}, \underline{V}_k^{\tilde{r}}]$, $G_k = \tilde{\underline{V}}_k^T \underline{V}_k$
  $p'_{k,0} = [1, 0_{1,2s}]^T$, $r'_{k,0} = [0_{1,s+1}, 1, 0_{1,s-1}]^T$, $x'_{k,0} = [0_{2s+1}]$
  $\tilde{p}'_{k,0} = [1, 0_{1,2s}]^T$, $\tilde{r}'_{k,0} = [0_{1,s+1}, 1, 0_{1,s-1}]^T$
  for $j = 0 : s-1$ **do**
    $\alpha_{sk+j} = ((\tilde{r}'_{k,j})^T G_k r'_{k,j})/((\tilde{p}'_{k,j})^T G_k \underline{B}_k p'_{k,j})$
    $x'_{k,j+1} = x'_{k,j} + \alpha_{sk+j} p'_{k,j}$
    $r'_{k,j+1} = r'_{k,j} - \underline{B}_k \left(\alpha_{sk+j} p'_{k,j}\right)$
    $\tilde{r}'_{k,j+1} = \tilde{r}'_{k,j} - \underline{B}_k \left(\alpha_{sk+j} \tilde{p}'_{k,j}\right)$
    $\beta_{sk+j+1} = ((\tilde{r}'_{k,j+1})^T G_k r'_{k,j+1})/((\tilde{r}'_{k,j})^T G_k r'_{k,j})$
    $p'_{k,j+1} = r'_{k,j+1} + \beta_{sk+j+1} p'_{k,j}$
    $\tilde{p}'_{k,j+1} = \tilde{r}'_{k,j+1} + \beta_{sk+j+1} \tilde{p}'_{k,j}$
  end for
  $x_{sk+s} = \underline{V}_k x'_{k,s} + x_{sk}$, $r_{sk+s} = \underline{V}_k r'_{k,s}$, $p_{sk+s} = \underline{V}_k p'_{k,s}$
  $\tilde{r}_{sk+s} = \tilde{\underline{V}}_k \tilde{r}'_{k,s}$, $\tilde{p}_{sk+s} = \tilde{\underline{V}}_k \tilde{p}'_{k,s}$
  $k = k + 1$
end while
return x_{sk}

with starting vector $v_{k,0}^p = p_{sk}$. We assume we use the same recurrence in constructing $v_{k,i}^r$. The choice of parameters γ_i, θ_i, and σ_i play a large role in determining the quality of the resulting basis, which in turn affects both stability and convergence in s-step BICG. We denote $\underline{V}_k = [\underline{V}_k^p, \underline{V}_k^r]$. We also denote $V_k = [V_k^p, 0_N, V_k^r, 0_N]$ where V_k^p and V_k^r are $\underline{V}_k^p$ and $\underline{V}_k^r$, resp., with their last columns omitted.

Within the inner loop, in step j of outer loop k, we update the length-$(2s+1)$ coefficients for the BICG vectors as linear combinations of the columns in $\underline{V}_k$ (rather than explicitly update the length-N BICG vectors, as in classical BICG). The coefficient vectors are denoted with prime symbols (i.e., $r_{sk+j} = \underline{V}_k r'_{k,j}$, and similarly for p_{sk+j} and x_{sk+j}). The inner iteration updates then become

$$r'_{k,j+1} = r'_{k,j} - \alpha_{sk+j} \underline{B}_k p'_{k,j} \quad \text{and} \tag{2.2}$$

$$p'_{k,j+1} = r'_{k,j+1} + \beta_{sk+j+1} p'_{k,j}, \tag{2.3}$$

where

$$\underline{B}_k = \begin{bmatrix} \begin{bmatrix} \underline{C}_{k,s+1} & 0_{s+1,1} \end{bmatrix} & \\ & \begin{bmatrix} \underline{C}_{k,s} & 0_{s,1} \end{bmatrix} \end{bmatrix},$$

with

$$\underline{C}_{k,j} = \begin{bmatrix} \theta_0 & -\sigma_1/\gamma_1 & & \\ 1/\gamma_0 & \theta_1 & \ddots & \\ & 1/\gamma_1 & \ddots & -\sigma_{j-1}/\gamma_{j-1} \\ & & \ddots & \theta_{j-1} \\ & & & 1/\gamma_{j-1} \end{bmatrix}.$$

We can rearrange (2.2) and (2.3) as

$$\underline{B}_k p'_{k,j} = \frac{1}{\alpha_{sk+j}} \left(r'_{k,j} - r'_{k,j+1} \right) \quad \text{and} \tag{2.4}$$

$$r'_{k,j} = p'_{k,j} - \beta_{sk+j} p'_{k,j-1}. \tag{2.5}$$

Premultiplying (2.5) by $\underline{V}_k$, we obtain

$$V_k r'_{k,j} = V_k p'_{k,j} - \beta_{sk+j} V_k p'_{k,j-1}. \tag{2.6}$$

This equation is valid for $1 \leq j < s$, since $p'_{k,-1}$ is undefined. When $j = 0$, we have

$$\begin{aligned} V_k r'_{k,0} &= \underline{V}_{k-1} r'_{k-1,s} \\ &= \underline{V}_{k-1} \left(p'_{k-1,s} - \beta_{sk} p'_{k-1,s-1} \right) \\ &= V_k p'_{k,0} - \beta_{sk} V_{k-1} p'_{k-1,s-1}, \end{aligned}$$

which gives a valid expression for the $j = 0$ case.

Now, let

$$R'_{k,j} = \left[r'_{k,0}, r'_{k,1}, \ldots, r'_{k,j} \right] \qquad \text{and} \quad P'_{k,j} = \left[p'_{k,0}, p'_{k,1}, \ldots, p'_{k,j} \right].$$

We can write (2.6) in block form as

$$V_k R'_{k,j} = V_k P'_{k,j} U_{k,j} - \beta_{sk} V_{k-1} p'_{k-1,s-1} e_1^T, \tag{2.7}$$

where

$$U_{k,j} = \begin{bmatrix} 1 & -\beta_{sk+1} & & \\ & 1 & \ddots & \\ & & \ddots & -\beta_{sk+j} \\ & & & 1 \end{bmatrix}.$$

Premultiplying (2.7) by A, we obtain

$$A V_k R'_{k,j} = A V_k P'_{k,j} U_{k,j} - \beta_{sk} A V_{k-1} p'_{k-1,s-1} e_1^T. \tag{2.8}$$

We can also write (2.4) in block form as

$$\underline{B}_k P'_{k,j} = R'_{k,j} L_{k,j} \Lambda_{k,j}^{-1} - \frac{1}{\alpha_{sk+j}} r'_{k,j+1} e_{j+1}^T, \tag{2.9}$$

where $\Lambda_{k,j} = \operatorname{diag}\left(\alpha_{sk}, \ldots, \alpha_{sk+j} \right)$ and

$$L_{k,j} = \begin{bmatrix} 1 & & & \\ -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{bmatrix}.$$

If we premultiply (2.9) by $\underline{V}_k$ and postmultiply by $U_{k,j}$, we obtain

$$\underline{V}_k \underline{B}_k P'_{k,j} U_{k,j} = \underline{V}_k R'_{k,j} L_{k,j} \Lambda_{k,j}^{-1} U_{k,j} - \frac{1}{\alpha_{sk+j}} \underline{V}_k r'_{k,j+1} e_{j+1}^T,$$

which can be written

$$AV_k P'_{k,j} U_{k,j} = V_k R'_{k,j} L_{k,j} \Lambda_{k,j}^{-1} U_{k,j} - \frac{1}{\alpha_{sk+j}} \underline{V}_k r'_{k,j+1} e_{j+1}^T \tag{2.10}$$

since $AV_k = \underline{V}_k \underline{B}_k$ and $\underline{V}_k R'_{k,j} = V_k R'_{k,j}$ for $j \leq s-1$. We can then combine (2.8) and (2.10) to obtain

$$AV_k R'_{k,j} = V_k R'_{k,j} \hat{T}_{k,j} - \frac{\beta_{sk}}{\alpha_{sk-1}} V_{k-1} r'_{k-1,s-1} e_1^T - \frac{1}{\alpha_{sk+j}} \underline{V}_k r'_{k,j+1} e_{j+1}^T, \tag{2.11}$$

where $\hat{T}_{k,j} = L_{k,j} \Lambda_{k,j}^{-1} U_{k,j} + e_1 \frac{\beta_{sk}}{\alpha_{sk-1}} e_1^T$. Note when $k = 0$, $\frac{\beta_{sk}}{\alpha_{sk-1}}$ is defined to be 0.

We can now combine outer loop iterations in block form to write the s-step BICG recurrence for iterations 0 through $sk + j$. Let $\mathcal{V}_k = [V_0, V_1, \ldots, V_k]$. Let

$$\mathcal{R}'_{k,j} = \begin{bmatrix} R'_{0,s-1} & & & \\ & R'_{1,s-1} & & \\ & & \ddots & \\ & & & R'_{k,j} \end{bmatrix}$$

and

$$\mathcal{T}_{k,j} = \begin{bmatrix} \frac{1}{\alpha_0} & -\frac{\beta_1}{\alpha_0} & & \\ -\frac{1}{\alpha_0} & \frac{1}{\alpha_1} + \frac{\beta_1}{\alpha_0} & \ddots & \\ & \ddots & \ddots & \frac{\beta_{sk+j}}{\alpha_{sk+j-1}} \\ & & -\frac{1}{\alpha_{sk+j-1}} & \frac{1}{\alpha_{sk+j}} + \frac{\beta_{sk+j}}{\alpha_{sk+j-1}} \end{bmatrix}.$$

Then by (2.11), we can write

$$A\mathcal{V}_k \mathcal{R}'_{k,j} = \mathcal{V}_k \mathcal{R}'_{k,j} \mathcal{T}_{k,j} - \frac{1}{\alpha_{sk+j}} \underline{V}_k r'_{k,j+1} e_{sk+j+1}^T.$$

Since we can write the residual vectors as $\mathcal{R}_n = [r_0, \ldots, r_n] = \mathcal{V}_k \mathcal{R}'_{k,j}$, where $n = sk + j$, we can write the above as

$$A\mathcal{R}_n = \mathcal{R}_n \mathcal{T}_n - \frac{1}{\alpha_n} r_{n+1} e_{n+1}^T,$$

which gives us the same governing equation for iterations 0 through $sk + j$ as the classical BICG algorithm in exact arithmetic [25]. Note that a similar relation holds for the dual Krylov vectors $\tilde{r}_{sk+j}$ and $\tilde{p}_{sk+j}$.

3. s-step BICG in finite precision. The goal of this section is to derive a Lanczos-type recurrence for finite precision s-step BICG of the form

$$A\mathcal{V}_k \mathcal{R}'_{k,j} = \mathcal{V}_k \mathcal{R}'_{k,j} \mathcal{T}_{k,j} - \frac{1}{\alpha_{sk+j}} \underline{V}_k r'_{k,j+1} e_{sk+j+1}^T + \epsilon \mathbf{\Delta}_{k,j}$$

and upper bound the size of the error term $\epsilon \mathbf{\Delta}_{k,j}$. We assume a standard model of floating point arithmetic, where

$$\begin{aligned} \mathrm{fl}(\alpha x + y) &= \alpha x + y + \delta_1, && \text{where} \quad |\delta_1| \leq \epsilon 2 |\alpha x| + |y| + O(\epsilon^2), \quad \text{and} \\ \mathrm{fl}(Ax) &= Ax + \delta_2, && \text{where} \quad |\delta_2| \leq \epsilon N |A| |x| + O(\epsilon^2), \end{aligned}$$

where $x, y \in \mathbb{R}^N$, $\alpha \in \mathbb{R}$. In the remaining analysis we drop higher powers of ϵ for simplicity. Let ϵ be the machine precision unit. For simplicity of notation, we now let $r'_{k,j}$, $p'_{k,j}$, $x'_{k,j}$, α_{sk+j}, $r_{k,s}$, $p_{k,s}$, β_{sk+j}, $\underline{V}_k$, and $\underline{B}_k$ be the computed quantities in finite precision s-step BICG.

At the $(sk+j)^{\text{th}}$ iteration, to compute $r'_{k,j+1}$ we first compute $\underline{B}_k p'_{k,j}$ and have

$$\text{fl}\left(\underline{B}_k p'_{k,j}\right) = \underline{B}_k p'_{k,j} + g, \quad \text{where} \quad |g| \leq \epsilon(2s+1)\left|\underline{B}_k\right|\left|p'_{k,j}\right|.$$

Then

$$\begin{aligned} r'_{k,j+1} &= \text{fl}\left(r'_{k,j} - \alpha_{sk+j} \cdot \text{fl}\left(\underline{B}_k p'_{k,j}\right)\right) \\ &= r'_{k,j} - \alpha_{sk+j}\underline{B}_k p'_{k,j} - \alpha_{sk+j} g + g', \end{aligned} \tag{3.1}$$

where $\left|g'\right| \leq \epsilon\left(\left|r'_{k,j}\right| + 2\left|\alpha_{sk+j}\right|\left|\underline{B}_k p'_{k,j}\right|\right)$. Let $\delta_{r'_{k,j}} = \left(\alpha_{sk+j} g + g'\right)/\left(\epsilon\left|\alpha_{sk+j}\right|\right)$. Then using (3.1) we obtain

$$\frac{1}{\alpha_{sk+j}}\left(r'_{k,j+1} - r'_{k,j}\right) = -\underline{B}_k p'_{k,j} + \epsilon\delta_{r'_{k,j}},$$

where

$$\left|\delta_{r'_{k,j}}\right| \leq (2s+1)\left|\underline{B}_k\right|\left|p'_{k,j}\right| + \frac{\left|r'_{k,j}\right|}{\left|\alpha_{sk+j}\right|} + 2\left|\underline{B}_k p'_{k,j}\right|. \tag{3.2}$$

Similarly,

$$\begin{aligned} p'_{k,j+1} &= \text{fl}\left(r'_{k,j+1} + \beta_{sk+j+1} p'_{k,j}\right) \\ &= r'_{k,j+1} + \beta_{sk+j+1} p'_{k,j} + f, \end{aligned}$$

where $|f| \leq \epsilon\left(\left|r'_{k,j+1}\right| + 2\left|\beta_{sk+j+1}\right|\left|p'_{k,j}\right|\right)$. Letting $\delta_{p'_{k,j+1}} = f/\epsilon$, we have

$$p'_{k,j+1} = r'_{k,j+1} + \beta_{sk+j+1} p'_{k,j} + \epsilon\delta_{p'_{k,j+1}},$$

where

$$\left|\delta_{p'_{k,j+1}}\right| \leq \left|r'_{k,j+1}\right| + 2\left|\beta_{sk+j+1}\right|\left|p'_{k,j}\right|. \tag{3.3}$$

Rearranging (3.2) and (3.3), we can write

$$\begin{aligned} \underline{B}_k p'_{k,j} &= \frac{1}{\alpha_{sk+j}}\left(r'_{k,j} - r'_{k,j+1}\right) + \epsilon\delta_{r'_{k,j}} \quad \text{and} \\ r'_{k,j} &= p'_{k,j} - \beta_{sk+j} p'_{k,j-1} + \epsilon\delta_{p'_{k,j}}, \end{aligned} \tag{3.4}$$

and premultiplying (3.4) by V_k gives

$$V_k r'_{k,j} = V_k p'_{k,j} - \beta_{sk+j} V_k p'_{k,j-1} + \epsilon V_k \delta_{p'_{k,j}}.$$

This equation is valid for $1 \leq j < s$, since $p'_{k,-1}$ is undefined. When $j = 0$, we have

$$\begin{aligned} V_k r'_{k,0} &= fl(\underline{V}_{k-1} r'_{k-1,s}) \\ &= \underline{V}_{k-1} r'_{k-1,s} + \epsilon\phi^r_{k-1} \quad \text{and} \\ V_k p'_{k,0} &= fl(\underline{V}_{k-1} p'_{k-1,s}) \\ &= \underline{V}_{k-1} p'_{k-1,s} + \epsilon\phi^p_{k-1}, \end{aligned}$$

where $\left|\phi^r_{k-1}\right| \leq (2s+1)\left|\underline{V}_{k-1}\right|\left|r'_{k-1,s}\right|$ and $\left|\phi^p_{k-1}\right| \leq (2s+1)\left|\underline{V}_{k-1}\right|\left|p'_{k-1,s}\right|$. Then for $j = 0$, we can write

$$\begin{aligned}
V_k r'_{k,0} &= \underline{V}_{k-1} r'_{k-1,s} + \epsilon\phi^r_{k-1} \\
&= \underline{V}_{k-1}\left(p'_{k-1,s} - \beta_{sk} p'_{k-1,s-1} + \epsilon\delta_{p'_{k-1,s}}\right) + \epsilon\phi^r_{k-1} \\
&= V_k p'_{k,0} - \epsilon\phi^p_{k-1} - \beta_{sk} V_{k-1} p'_{k-1,s-1} + \epsilon\underline{V}_{k-1}\delta_{p'_{k-1,s}} + \epsilon\phi^r_{k-1} \\
&= V_k p'_{k,0} - \beta_{sk} V_{k-1} p'_{k-1,s-1} + \epsilon\left(\underline{V}_{k-1}\delta_{p'_{k-1,s}} + \phi^r_{k-1} - \phi^p_{k-1}\right).
\end{aligned}$$

Now, let $\Delta_{R'_{k,j}} = \left[\delta_{r'_{k,0}}, \ldots, \delta_{r'_{k,j}}\right]$ and $\Delta_{P'_{k,j}} = \left[0_{2s+1}, \delta_{p'_{k,1}}, \ldots, \delta_{p'_{k,j}}\right]$. We can then write

$$\begin{aligned}
V_k R'_{k,j} =& V_k P'_{k,j} U_{k,j} - \beta_{sk} V_{k-1} p'_{k-1,s-1} e_1^T + \epsilon V_k \Delta_{P'_{k,j}} \\
&+ \epsilon\left(\underline{V}_{k-1}\delta_{p'_{k-1,s}} + \phi^r_{k-1} - \phi^p_{k-1}\right) e_1^T \quad \text{and} \qquad (3.5) \\
\underline{B}_k P'_{k,j} =& R'_{k,j} L_{k,j} \Lambda^{-1}_{k,j} - \frac{1}{\alpha_{sk+j}} r'_{k,j+1} e_{j+1}^T + \epsilon\Delta_{R'_{k,j}}. \qquad (3.6)
\end{aligned}$$

Premultiplying (3.5) by A gives

$$\begin{aligned}
AV_k R'_{k,j} =& AV_k P'_{k,j} U_{k,j} - \beta_{sk} AV_{k-1} p'_{k-1,s-1} e_1^T + \epsilon AV_k \Delta_{P'_{k,j}} \\
&+ \epsilon A\left(\underline{V}_{k-1}\delta_{p'_{k-1,s}} + \phi^r_{k-1} - \phi^p_{k-1}\right) e_1^T, \qquad (3.7)
\end{aligned}$$

and premuliplying (3.6) by $\underline{V}_k$ gives

$$\underline{V}_k \underline{B}_k P'_{k,j} = V_k R'_{k,j} L_{k,j} \Lambda^{-1}_{k,j} - \frac{1}{\alpha_{sk+j}} \underline{V}_k r'_{k,j+1} e_{j+1}^T + \epsilon V_k \Delta_{R'_{k,j}} \qquad (3.8)$$

for $j \leq s-1$.

Now, to write the error in s-step BICG in the context of classical BICG, we must account for error in computation of the s-step bases. Rearranging the finite precision evaluation of (2.1), we obtain

$$Av^p_{k,i} = \frac{1}{\gamma_i} v^p_{k,i+1} + \theta_i v^p_{k,i} - \frac{\sigma_i}{\gamma_i} v^p_{k,i-1} + \epsilon\delta_{v^p_{k,i}},$$

where we can write $\left|\delta_{v^p_{k,i}}\right|$ as

$$\left|\delta_{v^p_{k,i}}\right| = \big(N+2\big)\left|A\right|\left|v^p_{k,i}\right| + 3\left|\theta_i v^p_{k,i}\right| + \frac{3\left|\sigma_i v^p_{k,i-1}\right|}{\left|\gamma_i\right|}.$$

Since we generate $v^r_{k,i}$ by the same recurrence, we also have

$$\left|\delta_{v^r_{k,i}}\right| = \big(N+2\big)\left|A\right|\left|v^r_{k,i}\right| + 3\left|\theta_i v^r_{k,i}\right| + \frac{3\left|\sigma_i v^r_{k,i-1}\right|}{\left|\gamma_i\right|}.$$

Letting $\Delta_{V_k} = \left[\delta^p_{v_{k,0}}, \ldots, \delta^p_{v_{k,s-1}}, 0, \delta^r_{v_{k,0}}, \ldots, \delta^r_{v_{k,s-2}}, 0\right]$, we can then write the finite precision basis computation as

$$AV_k = \underline{V}_k \underline{B}_k + \epsilon\Delta_{V_k}. \qquad (3.9)$$

Using (3.9), we can write (3.8) as

$$(AV_k - \epsilon\Delta_{V_k})P'_{k,j} = V_k R'_{k,j} L_{k,j}\Lambda_{k,j}^{-1} - \frac{1}{\alpha_{sk+j}}\underline{V}_k r'_{k,j+1} e_{j+1}^T + \epsilon V_k \Delta_{R'_{k,j}},$$

which can be rearranged to obtain

$$AV_k P'_{k,j} = V_k R'_{k,j} L_{k,j}\Lambda_{k,j}^{-1} - \frac{1}{\alpha_{sk+j}}\underline{V}_k r'_{k,j+1} e_{j+1}^T + \epsilon V_k \Delta_{R'_{k,j}} + \epsilon\Delta_{V_k} P'_{k,j}. \tag{3.10}$$

Postmultiplying (3.10) by $U_{k,j}$ gives

$$\begin{aligned} AV_k P'_{k,j} U_{k,j} =& V_k R'_{k,j} L_{k,j}\Lambda_{k,j}^{-1} U_{k,j} - \frac{1}{\alpha_{sk+j}}\underline{V}_k r'_{k,j+1} e_{j+1}^T \\ &+ \epsilon\big(V_k \Delta_{R'_{k,j}} U_{k,j} + \Delta_{V_k} P'_{k,j} U_{k,j}\big), \end{aligned} \tag{3.11}$$

and combining (3.11) and (3.7), we obtain

$$AV_k R'_{k,j} = V_k R'_{k,j} L_{k,j}\Lambda_{k,j}^{-1} U_{k,j} - \frac{1}{\alpha_{sk+j}}\underline{V}_k r'_{k,j+1} e_{j+1}^T - \beta_{sk} AV_{k-1} p'_{k-1,s-1} e_1^T \tag{3.12}$$

$$+ \epsilon\left(AV_k \Delta_{P'_{k,j}} + V_k \Delta_{R'_{k,j}} U_{k,j} + \Delta_{V_k} P'_{k,j} U_{k,j}\right) \tag{3.13}$$

$$+ \epsilon A\left(\underline{V}_{k-1}\delta_{p'_{k-1,s}} + \phi_{k-1}^r - \phi_{k-1}^p\right) e_1^T. \tag{3.14}$$

Since

$$\begin{aligned} \beta_{sk} AV_{k-1} p'_{k-1,s-1} e_1^T =& \beta_{sk}(\underline{V}_{k-1}\underline{B}_{k-1} + \Delta_{V_{k-1}}) p'_{k-1,s-1} e_1^T \\ =& \beta_{sk}\underline{V}_{k-1}\left(\frac{1}{\alpha_{sk-1}}(r'_{k-1,s-1} - r'_{k-1,s}) + \epsilon\delta_{r'_{k-1,s-1}}\right) e_1^T \\ &+ \beta_{sk}\Delta_{V_{k-1}} p'_{k-1,s-1} e_1^T \\ =& \frac{\beta_{sk}}{\alpha_{sk-1}} V_{k-1} r'_{k-1,s-1} e_1^T - \frac{\beta_{sk}}{\alpha_{sk-1}}(V_k r'_{k,0} - \epsilon\phi_{k-1}^r) e_1^T \\ &+ \epsilon\beta_{sk} V_{k-1}\delta_{r'_{k-1,s-1}} e_1^T + \beta_{sk}\Delta_{V_{k-1}} p'_{k-1,s-1} e_1^T, \end{aligned}$$

we can write (3.14) as

$$AV_k R'_{k,j} = V_k R'_{k,j}\hat{T}_{k,j} - \frac{\beta_{sk}}{\alpha_{sk-1}} V_{k-1} r'_{k-1,s-1} e_1^T - \frac{1}{\alpha_{sk+j}}\underline{V}_k r'_{k,j+1} e_{j+1}^T + \epsilon\Delta_{k,j},$$

where

$$\Delta_{k,j} = AV_k \Delta_{P'_{k,j}} + AV_{k-1}\delta_{p'_{k-1,s}} e_1^T + V_k \Delta_{R'_{k,j}} U_{k,j} - \beta_{sk} V_{k-1}\delta_{r'_{k-1,s-1}} e_1^T \tag{3.15}$$

$$+ \Delta_{V_k} P'_{k,j} U_{k,j} - \beta_{sk}\Delta_{V_{k-1}} p'_{k-1,s-1} e_1^T + \left(A(\phi_{k-1}^r - \phi_{k-1}^p) - \frac{\beta_{sk}}{\alpha_{sk-1}}\phi_{k-1}^r\right) e_1^T. \tag{3.16}$$

Writing $\Delta_{k,j} = [\delta_{sk}, \ldots, \delta_{sk+j}]$, we have that the $(sk+j+1)^{\text{th}}$ column of $\Delta_{k,j}$ is

$$\delta_{sk+j} = AV_k \delta_{p'_{k,j}} + V_k \delta_{r'_{k,j}} - \beta_{sk+j} V_k \delta_{r'_{k,j-1}} + \Delta_{V_k} r'_{k,j-1}, \tag{3.17}$$

for $j > 0$, and

$$\begin{aligned}\delta_{sk} =& AV_{k-1}\delta_{p'_{k-1,s}} + V_k\delta_{r'_{k,0}} - \beta_{sk}V_{k-1}\delta_{r'_{k-1,s-1}} + \delta^p_{v_k,0} - \beta_{sk+j}\Delta_{V_{k-1}}p'_{k-1,s-1} \\ &+ \left(A(\phi^r_{k-1} - \phi^p_{k-1}) - \frac{\beta_{sk}}{\alpha_{sk-1}}\phi^r_{k-1}\right) \end{aligned} \tag{3.18}$$

for $j = 0$.

Using the inequalities $|\beta_{sk+j}p'_{k,j-1}| \leq |p'_{k,j}| + |r'_{k,j}| + O(\epsilon)$ and $|r'_{k,j-1}| \leq |r'_{k,j}| + |\alpha_{sk+j-1}||\underline{B}_k p'_{k,j-1}| + O(\epsilon)$, we can bound the norm of the columns as

$$\begin{aligned}|\delta_{sk+j}| \leq& \left((N+6)|A||V_k| + (2s+8)|\underline{V}_k||\underline{B}_k| + \left(\frac{1}{|\alpha_{sk+j}|} + \frac{|\beta_{sk+j}|}{|\alpha_{sk+j-1}|}\right)|V_k|\right)|r'_{k,j}| \\ &+ \left(2|A||V_k| + (4s+7)|\underline{V}_k||\underline{B}_k|\right)|p'_{k,j}|, \end{aligned} \tag{3.19}$$

if $j > 0$. For the $j = 0$ case, we have

$$|\delta_{sk}| \leq \left((N+2s+7)|A||\underline{V}_{k-1}| + (2s+8)|\underline{V}_{k-1}||\underline{B}_{k-1}|\right)|r'_{k-1,s}| \tag{3.20}$$

$$+ \left(\frac{1}{|\alpha_{sk}|} + \frac{(2s+2)|\beta_{sk}|}{|\alpha_{sk-1}|}\right)|\underline{V}_{k-1}||r'_{k-1,s}|$$

$$+ \left((2N+4s+16)|A||\underline{V}_{k-1}| + (6s+22)|\underline{V}_{k-1}||\underline{B}_{k-1}|\right)|p'_{k-1,s}| \tag{3.21}$$

We can thus write the finite precision s-step BICG recurrence for iterations 0 through $sk + j$ as

$$A\mathcal{V}_k\mathcal{R}'_{k,j} = \mathcal{V}_k\mathcal{R}'_{k,j}\mathcal{T}_{k,j} - \frac{1}{\alpha_{sk+j}}\underline{V}_k r'_{k,j+1}e^T_{sk+j+1} + \epsilon\mathbf{\Delta}_{k,j}, \tag{3.22}$$

where $\mathbf{\Delta}_{k,j} = [\Delta_{0,s-1}, \Delta_{1,s-1}, \ldots, \Delta_{k,j}]$.

3.1. Comments. Note that we can write n iterations of finite precision classical BICG as n iterations of finite precision s-step BICG with $s > n$, performed in the standard basis. That is, we have a single outer loop iteration $k = 0$ and $j = n$ inner loop iterations, with $V_0 = I_n$, $B_0 = A$, $R'_{0,n} = R_{0,n}$, and $P'_{0,n} = P_{0,n}$. Now, since $V_0 = I_n$, $\Delta_{V_0} = 0$, and since $k = 0$, ϕ^r_{k-1},ϕ^p_{k-1}, and $\delta_{p_{k-1,s}}$ are defined to be zero. Plugging in to (3.15), we get

$$\Delta_{0,n} = A\Delta_{P_{0,n}} + \Delta_{R_{0,n}}U_{0,n},$$

which reproduces the error term (modulo a factor of 2) obtained by Tong and Ye for finite precision classical BICG [25].

Also note that from (3.15), we can see that the first four terms on the right-hand side correspond to the two terms in Tong and Ye's analysis for classical BICG, and the remaining terms correspond to the error in computing the s-step Krylov bases and the change of basis operation. We can also see that a bound on the size of the error in each column of the finite precision recurrence depends on both the magnitude of the error in computing the s-step Krylov bases, i.e., $\|\Delta_{V_k}\|$, as well as the size of the bases, i.e., $\|V_k\|$.

3.2. Diagonal scaling. As in [25], it will be more convenient to work with a scaled version of (3.22) in subsequent sections. Let $Z_{k,j} = [z_{0,0}, \ldots, z_{k,j}] = \mathcal{V}_k \mathcal{R}'_{k,j} D_{k,j}^{-1}$ where

$$D_{k,j} = \operatorname{diag}\left(\left\|V_0 r'_{0,0}\right\|, \ldots, \left\|V_0 r'_{0,s-1}\right\|, \left\|V_1 r'_{1,0}\right\|, \ldots, \left\|V_k r'_{k,j}\right\|\right).$$

We can then write the scaled version of (3.22) as

$$AZ_{k,j} = Z_{k,j}\bar{\mathcal{T}}_{k,j} - \frac{1}{\bar{\alpha}_{sk+j}} \frac{\underline{V}_k r'_{k,j+1}}{\|r_0\|} e_{sk+j+1}^T + \epsilon \bar{\mathbf{\Delta}}_{k,j}, \tag{3.23}$$

where $\bar{\mathcal{T}}_{k,j} = D_{k,j}\mathcal{T}_{k,j}D_{k,j}^{-1}$,

$$\bar{\alpha}_{sk+j} = \left\|V_k r'_{k,j}\right\| \alpha_{sk+j} / \|r_0\| = \left\|V_k r'_{k,j}\right\| \alpha_{sk+j} / \left\|V_0 r'_{0,0}\right\| = e_{sk+j+1}^T \bar{\mathcal{T}}_{k,j}^{-1} e_1,$$

and

$$\bar{\mathbf{\Delta}}_{k,j} = \mathbf{\Delta}_{k,j} D_{k,j}^{-1}.$$

4. Bounds on $\|r_{sk+j+1}\|$ for finite precision s-step BICG. In this subsection, we upper bound the norm of the updated residual computed in iteration $sk+j$ of s-step BICG. First, we will review a series of Lemmas proved by Tong and Ye [25]. The proofs shown below are nearly identical to those given by Tong and Ye [25], although we have changed the notation and indexing for consistency with our s-step formulation[1].

LEMMA 4.1. *Assume*

$$AZ_{k,j} = Z_{k,j}\bar{\mathcal{T}}_{k,j} - \frac{1}{\bar{\alpha}_{sk+j}} \frac{\underline{V}_k r'_{k,j+1}}{\|r_0\|} e_{sk+j+1}^T$$

with $r_0 = \|r_0\| z_0$. *Then for any polynomial* $\rho(x) = \sum_{i=0}^{sk+j+1} \psi_i x^i$ *of degree* $\leq sk+j+1$,

$$\rho(A)z_0 = Z_{k,j}\rho(\bar{\mathcal{T}}_{k,j})e_1 + c_{sk+j}\underline{V}_k r'_{k,j+1},$$

where $c_{sk+j} = (-1)^{sk+j+1}\psi_{sk+j+1}/(\alpha_0 \cdots \alpha_{sk+j} \|r_0\|)$.

Proof. First, we will prove by induction that for $1 \leq i \leq sk+j$

$$A^i Z_{k,j} e_1 = Z_{k,j}\bar{\mathcal{T}}_{k,j}^i e_1. \tag{4.1}$$

For $i = 1$, we have

$$AZ_{k,j}e_1 = \left(Z_{k,j}\bar{\mathcal{T}}_{k,j} - \frac{1}{\bar{\alpha}_{sk+j}} \frac{\underline{V}_k r'_{k,j+1}}{\|r_0\|} e_{sk+j+1}^T\right) e_1 = Z_{k,j}\bar{\mathcal{T}}_{k,j}e_1.$$

Now, assume (4.1) holds for some $i < sk+j$. Then

$$\begin{aligned}
A^{i+1}Z_{k,j}e_1 &= A(A^i Z_{k,j}e_1) \\
&= A(Z_{k,j}\bar{\mathcal{T}}_{k,j}e_1) \\
&= \left(Z_{k,j}\bar{\mathcal{T}}_{k,j} - \frac{1}{\bar{\alpha}_{sk+j}} \frac{\underline{V}_k r'_{k,j+1}}{\|r_0\|} e_{sk+j+1}^T\right) \bar{\mathcal{T}}_{k,j}^i e_1 \\
&= Z_{k,j}\bar{\mathcal{T}}_{k,j}\bar{\mathcal{T}}_{k,j}^i e_1 = Z_{k,j}\bar{\mathcal{T}}_{k,j}^{i+1} e_1,
\end{aligned}$$

[1] One lemma presented is slightly different than what appears in [25] due to a minor mathematical error that we correct.

where we have used the fact that $e_{sk+j+1}^T \bar{\mathcal{T}}_{k,j}^i e_1 = 0$ when $i < sk+j$. Therefore the inductive hypothesis holds. Now consider the case $i = sk+j$. We then have

$$\begin{aligned} A^{sk+j+1} Z_{k,j} e_1 &= A(Z_{k,j} \bar{\mathcal{T}}_{k,j}^{sk+j} e_1) \\ &= \left(Z_{k,j} \bar{\mathcal{T}}_{k,j} - \frac{1}{\bar{\alpha}_{sk+j}} \frac{\underline{V}_k r'_{k,j+1}}{\|r_0\|} e_{sk+j+1}^T \right) \left(\hat{\underline{T}}_{k,j}^{sk+j} e_1 \right). \end{aligned}$$

Since it can be shown that $e_{sk+j+1}^T \bar{\mathcal{T}}_{k,j}^{sk+j} e_1 = (-1)^{sk+j} \left\| V_k r'_{k,j} \right\| \left(\alpha_0 \cdots \alpha_{sk+j} \|r_0\| \right)^{-1}$ and $\bar{\alpha}_{sk+j} = \left\| V_k r'_{k,j} \right\| \alpha_{sk+j} / \|r_0\|$, we have

$$\begin{aligned} A^{sk+j+1} Z_{k,j} e_1 &= Z_{k,j} \bar{\mathcal{T}}_{k,j} \bar{\mathcal{T}}_{k,j}^{sk+j} e_1 \\ &= Z_{k,j} \bar{\mathcal{T}}_{k,j}^{sk+j+1} e_1 + \frac{(-1)^{sk+j+1}}{\alpha_0 \cdots \alpha_{sk+j} \|r_0\|} \underline{V}_k r'_{k,j+1}. \end{aligned}$$

The lemma follows. □

We now use this result in proving the following identity.

LEMMA 4.2. *Assume*

$$A Z_{k,j} = Z_{k,j} \bar{\mathcal{T}}_{k,j} - \frac{1}{\bar{\alpha}_{sk+j}} \frac{\underline{V}_k r'_{k,j+1}}{\|r_0\|} e_{sk+j+1}^T$$

with $r_0 = \|r_0\| z_0$ *and* $\bar{\alpha}_{sk+j} = e_{sk+j+1}^T \bar{\mathcal{T}}_{k,j}^{-1} e_1$. *Assume that* $W^T \in \mathbb{R}^{(sk+j+1) \times N}$ *is a matrix such that* $W^T Z_{k,j} = I$ *and* $W^T \underline{V}_k r'_{k,j+1} = 0_{sk+j+1}$. *Then for any polynomial* $\rho(x)$ *of degree not exceeding* $sk+j$ *with* $\rho(0) = 1$, *we have*

$$\underline{V}_k r'_{k,j+1} = \left(I - A Z_{k,j} \bar{\mathcal{T}}_{k,j}^{-1} W^T \right) \rho(A) r_0.$$

Proof. First, we multiply by $\bar{\mathcal{T}}_{k,j}^{-1} e_1$ to get

$$A Z_{k,j} \bar{\mathcal{T}}_{k,j}^{-1} e_1 = \left(Z_{k,j} \bar{\mathcal{T}}_{k,j} - \frac{1}{\bar{\alpha}_{sk+j}} \frac{\underline{V}_k r'_{k,j+1}}{\|r_0\|} e_{sk+j+1}^T \right) \bar{\mathcal{T}}_{k,j}^{-1} e_1,$$

which allows us to write

$$\frac{\underline{V}_k r'_{k,j+1}}{\|r_0\|} = z_0 - A Z_{k,j} \bar{\mathcal{T}}_{k,j}^{-1} e_1.$$

Now, let $\rho(x) = 1 + x\phi(x)$, with $\phi(x) = \sum_{i=0}^{sk+j} \psi_{i+1} x^i$ a polynomial of degree not exceeding $sk+j$. Then

$$\begin{aligned} \frac{\underline{V}_k r'_{k,j+1}}{\|r_0\|} &= z_0 - A Z_{k,j} \bar{\mathcal{T}}_{k,j}^{-1} e_1 + (\rho(A) z_0 - \rho(A) z_0) \\ &= -A\phi(A) z_0 - A Z_{k,j} \bar{\mathcal{T}}_{k,j}^{-1} e_1 + \rho(A) z_0 \\ &= -A Z_{k,j} \phi(\bar{\mathcal{T}}_{k,j}) e_1 - A Z_{k,j} \bar{\mathcal{T}}_{k,j}^{-1} e_1 + \rho(A) z_0 \\ &= -A Z_{k,j} (\phi(\bar{\mathcal{T}}_{k,j}) + \bar{\mathcal{T}}_{k,j}^{-1}) e_1 + \rho(A) z_0 \\ &= -A Z_{k,j} \bar{\mathcal{T}}_{k,j}^{-1} \rho(\bar{\mathcal{T}}_{k,j}) e_1 + \rho(A) z_0. \end{aligned} \tag{4.2}$$

By Lemma 4.1, recall that

$$\rho(A)z_0 = Z_{k,j}\rho(\bar{\mathcal{T}}_{k,j})e_1 + c_{sk+j}\underline{V}_k r'_{k,j+1},$$

and, multiplying by W^T, we have

$$\begin{aligned} W^T\rho(A)z_0 &= W^T\left(Z_{k,j}\rho(\bar{\mathcal{T}}_{k,j})e_1 + c_{sk+j}\underline{V}_k r'_{k,j+1}\right) \\ &= \rho(\bar{\mathcal{T}}_{k,j})e_1, \end{aligned}$$

since $W^T Z_{k,j} = I$ and $W^T\underline{V}_k r'_{k,j+1} = 0_{sk+j+1}$. Now, we can write

$$\begin{aligned} \frac{\underline{V}_k r'_{k,j+1}}{\|r_0\|} &= -AZ_{k,j}\bar{\mathcal{T}}_{k,j}^{-1}W^T\rho(A)z_0 + \rho(A)z_0 \\ &= \left(I - AZ_{k,j}\bar{\mathcal{T}}_{k,j}^{-1}W^T\right)\rho(A)z_0, \end{aligned}$$

and substituting $z_0 = r_0/\|r_0\|$, we obtain

$$\underline{V}_k r'_{k,j+1} = \left(I - AZ_{k,j}\bar{\mathcal{T}}_{k,j}^{-1}W^T\right)\rho(A)r_0,$$

which gives the desired result. ∎

The following lemma describes the construction of the basis W.

LEMMA 4.3. *Assume that $z_0, \ldots, z_{sk+j+1} \in \mathbb{R}^N$ are linearly independent and write $Z_{k,j} = [z_0, \ldots, z_{sk+j}]$, $\underline{Z}_{k,j} = [Z_{k,j}, z_{sk+j+1}]$. Then $W_0^T = [I_{sk+j+1}, 0_{sk+j+1}]\underline{Z}_{k,j}^+$ has the property $W_0^T Z_{k,j} = I$ and $W_0^T z_{sk+j+1} = 0_{sk+j+1}$. Furthermore, its spectral norm is minimal among all matrices having this property.*

Proof. By the definition of W_0, $\underline{Z}_{k,j}^+ = [W_0, w]^T$ for some w. Since we assume $z_0, \ldots, z_{sk+j+1}$ are linearly independent,

$$[W_0, w]^T[Z_{k,j}, z_{sk+j+1}] = \underline{Z}_{k,j}^+\underline{Z}_{k,j} = I.$$

Then $W_0^T Z_{k,j} = I_{sk+j+1}$ and $W_0^T z_{sk+j+1} = 0_{sk+j+1}$.

Now, assume W is some other matrix such that $W^T Z_{k,j} = I$ and $W^T z_{sk+j+1} = 0_{sk+j+1}$ hold. Then $W^T[Z_{k,j}, z_{sk+j+1}] = [I_{sk+j+1}, 0_{sk+j+1}]$. Thus, $W^T\underline{Z}_{k,j}\underline{Z}_{k,j}^+ = [I_{sk+j+1}, 0_{sk+j+1}]\underline{Z}_{k,j}^+ = W_0^T$. Hence $\|W_0\| \le \|W\| \cdot \left\|\underline{Z}_{k,j}\underline{Z}_{k,j}^+\right\| \le \|W\|$. ∎

We can now present the main result.

THEOREM 4.4. *Assume* (3.23) *holds and let $W_0^T = [I_{sk+j+1}, 0_{sk+j+1}]\underline{Z}_{k,j}^+ \in \mathbb{R}^{(sk+j+1)\times N}$. If $z_0, \ldots, z_{sk+j+1}$ are linearly independent, then*

$$\left\|\underline{V}_k r'_{k,j+1}\right\| \le (1 + K_{k,j}) \min_{\rho\in\mathbb{P}_{sk+j+1}, \rho(0)=1} \|\rho(A + \delta A_{k,j})r_0\|, \tag{4.3}$$

where $K_{k,j} = \left\|(AZ_{k,j} - \epsilon\bar{\mathbf{\Delta}}_{k,j})\bar{\mathcal{T}}_{k,j}^{-1}W_0^T\right\|$ and $\delta A_{k,j} = -\epsilon\mathbf{\Delta}_{k,j}Z_{k,j}^+$.

Proof. Since $z_0, \ldots, z_{sk+j+1}$ are linearly independent, $Z_{k,j}^+ Z_{k,j} = I$. Then $\delta A_{k,j} = -\epsilon\bar{\mathbf{\Delta}}_{k,j}Z_{k,j}^+ \in \mathbb{R}^{N\times N}$ satisfies $\delta A_{k,j}Z_{k,j} = -\epsilon\bar{\mathbf{\Delta}}_{k,j}$. Thus (3.23) can be written as

$$(A + \delta A_{k,j})Z_{k,j} = Z_{k,j}\bar{\mathcal{T}}_{k,j} - \frac{1}{\bar{\alpha}_{sk+j}}\frac{\underline{V}_k r'_{k,j+1}}{\|r_0\|}e_{sk+j+1}^T. \tag{4.4}$$

Then, by Lemma 4.2, for any $\rho \in \mathbb{P}_{sk+j+1}$ with $\rho(0) = 1$, we obtain

$$\begin{aligned}\underline{V}_k r'_{k,j+1} &= (I - (A + \delta A_{k,j}) Z_{k,j} \bar{\mathcal{T}}_{k,j}^{-1} W_0^T) \cdot \rho(A + \delta A_{k,j}) r_0 \\ &= (I - (AZ_{k,j} - \epsilon \bar{\mathbf{\Delta}}_{k,j}) \bar{\mathcal{T}}_{k,j}^{-1} W_0^T) \cdot \rho(A + \delta A_{k,j}) r_0.\end{aligned}$$

Thus, we can bound the norm of the left hand side by

$$\left\|\underline{V}_k r'_{k,j+1}\right\| \leq \left(1 + \left\|(AZ_{k,j} - \epsilon \bar{\mathbf{\Delta}}_{k,j}) \bar{\mathcal{T}}_{k,j}^{-1} W_0^T\right\|\right) \cdot \left\|\rho(A + \delta A_{k,j}) r_0\right\|.$$

Since this holds for any $\rho(x)$ with $\rho(0) = 1$, the inequality is true for the minimizing polynomial, which leads to the bound. ∎

Note that $\tau_{k,j} = \min_{\rho \in \mathbb{P}_{sk+j+1}, \rho(0)=1} \|\rho(A + \delta A_{k,j}) r_0\|$ is the $(sk+j)^{\text{th}}$ residual norm of exact GMRES applied to the perturbed matrix $A + \delta A_{k,j}$, which decreases monotonically with increasing $(sk + j)$.

Since we have $K_{k,j} = \left\|(AZ_{k,j} - \epsilon \bar{\mathbf{\Delta}}_{k,j}) \bar{\mathcal{T}}_{k,j}^{-1} W_0^T\right\|$, we can bound $K_{k,j}$ as

$$K_{k,j} \leq \left(\sqrt{sk + j + 1}\|A\| + \epsilon \left\|\bar{\mathbf{\Delta}}_{k,j}\right\|\right) \left\|\bar{\mathcal{T}}_{k,j}^{-1}\right\| \cdot \|W_0\|.$$

Then, assuming $\left\|\bar{\mathcal{T}}_{k,j}^{-1}\right\|$ and $\|W_0\|$ are bounded, $\left\|\underline{V}_k r'_{k,j+1}\right\|$ is on the order $O(\tau_{k,j})$. We therefore expect convergence of the s-step BICG residual when $K_{k,j}$ increases at a slower rate than $\tau_{k,j}$ decreases, for all values of k.

Unfortunately, as in the BICG case, we can not determine $K_{k,j}$ *a priori*, although we can make some meaningful observations based on the bound in (4.3). Clearly, the terms $\epsilon \bar{\mathbf{\Delta}}_{k,j}$ in $K_{k,j}$ and

Note that in the case of CG (SPD A), we have $\left\|\underline{V}_k r'_{k,j+1}\right\|_2 = \left\|r_{sk+j+1}\right\|_2 = \left\|e^*_{sk+j+1}\right\|_A$, where e^*_{sk+j} denotes the solution error $e^*_{sk+j} = x^* - x_{sk+j}$ for true solution x^*. Thus in this case Theorem 4.4 gives a bound on the error of finite precision s-step CG. It remains future work to determine under what conditions $\left\|e^*_{sk+j+1}\right\|_A < \left\|e^*_{sk+j}\right\|_A$ for s-step CG.

5. The linearly dependent case. In the analysis above, we assumed linear independence among the residual vectors (which are scalar multiples of the Lanczos vectors). For many linear systems, however, convergence of classical BICG in finite precision is still observed despite numerical rank deficiency of the basis. In [25] it is shown how the residual norm can be bounded absent the assumption of linear independence, which gives insight into why convergence still occurs in such cases. We will now prove similar bounds, relaxing the constraint that $z_0, \ldots, z_{sk+j+1} \in \mathbb{R}^N$ be linearly independent. Again, our analysis extends that of Tong and Ye [25] for classical BICG.

We note that in the s-step case, there are two potential causes of a rank-deficient basis. Since we have $\mathcal{R}_{k,j} = \mathcal{V}_{k,j} \mathcal{R}'_{k,j}$, linear dependence can occur as a result of the finite precision Lanczos process, as in the classical method, as well as from numerical rank deficiencies in the generated s-step polynomial bases V_k.

Given $A \in \mathbb{R}^{N \times N}$ and $B \in \mathbb{R}^{N' \times N'}$, $AE - EB = Z$ corresponds to the linear system with coefficient matrix $A \otimes I_{N'} - I_N \otimes B$. This system has a unique solution if and only if $\lambda(A) \cap \lambda(B) = \emptyset$, or, equivalently, if $\text{sep}(A, B) := \left\|(A \otimes I_{N'} - I_N \otimes B)^{-1}\right\|^{-1} > 0$, which depends on the spectral gap of A and B (see [9]).

THEOREM 5.1. *Assume* (3.23) *holds, and let* μ *be a complex number such that* $\text{sep}(A - \mu I, \bar{\mathcal{T}}_{k,j}) \gg 0$. *Then*

$$\left\|\underline{V}_k r'_{k,j+1}\right\| \leq K_{k,j} \min_{\rho \in \mathbb{P}_{sk+j+1}, \rho(0)=1} \left(\left\|\rho(\bar{\mathcal{T}}_{k,j})\right\| + \|\rho(A - \mu I)\|\right) \|r_0\|,$$

where

$$K_{k,j} = \frac{\sqrt{sk+j+1}(\mathrm{sep}(A-\mu I, \bar{\mathcal{T}}_{k,j}) + |\mu|) + \epsilon \left\| \bar{\mathbf{\Delta}}_{k,j} \right\|_F}{\mathrm{sep}(A-\mu I, \bar{\mathcal{T}}_{k,j})} \cdot \max\left(1, \|\rho(A-\mu I)\| \cdot \left\|\rho(\bar{\mathcal{T}}_{k,j})\right\|\right).$$

Proof. By (3.23),

$$(A-\mu I)Z_{k,j} = Z_{k,j}\bar{\mathcal{T}}_{k,j} - \frac{1}{\bar{\alpha}_{sk+j}} \frac{\underline{V}_k r'_{k,j+1}}{\|r_0\|} e_{sk+j+1}^T + \epsilon \bar{\mathbf{\Delta}}_{k,j} - \mu Z_{k,j}. \tag{5.1}$$

Then since $\mathrm{sep}(A-\mu I, \bar{\mathcal{T}}_{k,j}) > 0$, the equation

$$(A-\mu I)E_{k,j} = E_{k,j}\bar{\mathcal{T}}_{k,j} - \epsilon \bar{\mathbf{\Delta}}_{k,j} + \mu Z_{k,j} \tag{5.2}$$

has a unique solution $E_{k,j}$ with

$$\|E_{k,j}\|_F \leq \frac{\left\| -\epsilon \bar{\mathbf{\Delta}}_{k,j} + \mu Z_{k,j} \right\|_F}{\mathrm{sep}(A-\mu I, \bar{\mathcal{T}}_{k,j})} \leq \frac{\epsilon \left\| \bar{\mathbf{\Delta}}_{k,j} \right\|_F + |\mu| \sqrt{sk+j+1}}{\mathrm{sep}(A-\mu I, \bar{\mathcal{T}}_{k,j})}.$$

Combining (5.1) and (5.2), we can write

$$(A-\mu I)(Z_{k,j}+E_{k,j}) = (Z_{k,j}+E_{k,j})\bar{\mathcal{T}}_{k,j} - \frac{1}{\bar{\alpha}_{sk+j}} \frac{\underline{V}_k r'_{k,j+1}}{\|r_0\|} e_{sk+j+1}^T.$$

Thus, for any $\rho \in \mathbb{P}_{sk+j+1}$, $\rho(0) = 1$, we have, by (4.2),

$$\frac{\underline{V}_k r'_{k,j+1}}{\|r_0\|} = \rho(A-\mu I)(Z_{k,j}+E_{k,j})e_1 - (A-\mu I)(Z_{k,j}+E_{k,j})\bar{\mathcal{T}}_{k,j}^{-1}\rho(\bar{\mathcal{T}}_{k,j})e_1,$$

and thus

$$\begin{aligned} \frac{\left\| \underline{V}_k r'_{k,j+1} \right\|}{\|r_0\|} \leq & \left(\|Z_{k,j}\| + \|E_{k,j}\|\right) \|\rho(A-\mu I)\| \\ & + \|A-\mu I\| \left(\|Z_{k,j}\| + \|E_{k,j}\|\right) \left\| \bar{\mathcal{T}}_{k,j}^{-1} \right\| \left\| \rho(\bar{\mathcal{T}}_{k,j}) \right\|. \end{aligned}$$

Since

$$\|Z_{k,j}\| + \|E_{k,j}\| \leq \sqrt{sk+j+1} + \frac{\epsilon \left\| \bar{\mathbf{\Delta}}_{k,j} \right\|_F + |\mu| \sqrt{sk+j+1}}{\mathrm{sep}(A-\mu I, \bar{\mathcal{T}}_{k,j})},$$

we obtain the desired result. □

Note that in this case, if μ is such that $\mathrm{sep}(A-\mu I, \bar{\mathcal{T}}_{k,j})$ is large, the quantity $K_{k,j}$ depends heavily on $\left\| \bar{\mathcal{T}}_{k,j}^{-1} \right\|$. The minimizing polynomial part of the bound now depends on both $\rho(\bar{\mathcal{T}}_{k,j})$ and $\rho(A-\mu I)$.

6. Extensions: perturbation theory. We can think of (4.4) as an exact subspace relation for a perturbed A, i.e., the quantities $\mathcal{V}_k$, $\mathcal{R}'_{k,j}$, and $\mathcal{T}_{k,j}$ produced by the finite precision s-step BICG algorithm satisfy an exact subspace recurrence (4.4) for the perturbed system $A + \delta A_{k,j}$. This means that the eigenvalues of the computed matrix $\mathcal{T}_{k,j}$ generated by the s-step algorithm are among the eigenvalues of the

perturbed matrix $A - \epsilon \mathbf{\Delta}_{k,j} \mathcal{R}_{k,j}^{+} \mathcal{V}_k^{+}$. In the next theorem, we bound the distance of these eigenvalues to eigenvalues of unperturbed matrix A.

THEOREM 6.1. *Let A be a normal $n \times n$ matrix of full rank. For each eigenvalue μ of the matrix $\mathcal{T}_{k,j}$ computed by the finite precision s-step (BI)CG method, there exists an eigenvalue λ of A such that*

$$\left|\gamma - \mu\right| \leq \epsilon \left\| \mathbf{\Delta}_{k,j} \right\|_2 \left\| \mathcal{R}_{k,j}'^{+} \right\|_2 \left\| \mathcal{V}_k^{+} \right\|_2 \tag{6.1}$$

Proof. Note that $\bar{\mathcal{T}}_{k,j} = D_{k,j} \mathcal{T}_{k,j} D_{k,j}^{-1}$ has the same eigenvalues as $\mathcal{T}_{k,j}$. By application of the Bauer-Fike theorem [2] to (4.4), there exists an eigenvalue of γ of A such that

$$\left|\gamma - \mu\right| \leq \epsilon \left\| \bar{\mathbf{\Delta}}_{k,j} Z_{k,j}^{+} \right\|_2. \tag{6.2}$$

We can then write

$$\begin{aligned} \left\| \bar{\mathbf{\Delta}}_{k,j} Z_{k,j}^{+} \right\|_2 =& \left\| \mathbf{\Delta}_{k,j} D_{k,j}^{-1} D_{k,j} \mathcal{R}_{k,j}'^{+} \mathcal{V}_k^{+} \right\|_2 & (6.3) \\ \leq& \left\| \mathbf{\Delta}_{k,j} \right\|_2 \left\| \mathcal{R}_{k,j}'^{+} \right\|_2 \left\| \mathcal{V}_k^{+} \right\|_2 & (6.4) \end{aligned}$$

□

The right hand side above can be shown to depend on $\kappa(V_k)$ and $\kappa(R'_{k,j})$. The above theorem means that the Lanczos vectors computed by the s-step (BI)CG algorithm, $\mathcal{V}_k \mathcal{R}'_{k,j}$, span Krylov spaces of a matrix within $\epsilon \left\| \mathbf{\Delta}_{k,j} \mathcal{R}_{k,j}^{+} \mathcal{V}_k^{+} \right\|$ of A. Similar observations have been made for classical finite precision Krylov methods [21, 29].

In [21], Paige shows that for classical Lanczos without reorthogonalization, the perturbed matrix is very close to A until a Ritz value has stabilized. It is an open question whether a similar result (perhaps with additional restrictions on $\mathcal{V}_k$) applies to the s-step case.

REFERENCES

[1] Z. BAI, D. HU, AND L. REICHEL, *A Newton basis GMRES implementation*, IMA J. Numer. Anal., 14 (1994), pp. 563–581.

[2] F. BAUER AND C. FIKE, *Norms and exclusion theorems*, Numerische Mathematik, 2 (1960), pp. 137–141.

[3] E. CARSON, N. KNIGHT, AND J. DEMMEL, *Avoiding communication in nonsymmetric Lanczos-based Krylov subspace methods*, SIAM J. Sci. Comp., 35 (2013).

[4] A. CHRONOPOULOS AND C. GEAR, *On the efficient implementation of preconditioned s-step conjugate gradient methods on multiprocessors with memory hierarchy*, Parallel Comput., 11 (1989), pp. 37–53.

[5] ———, *s-step iterative methods for symmetric linear systems*, J. Comput. Appl. Math, 25 (1989), pp. 153–168.

[6] A. CHRONOPOULOS AND C. SWANSON, *Parallel iterative s-step methods for unsymmetric linear systems*, Parallel Comput., 22 (1996), pp. 623–641.

[7] J. DEMMEL, M. HOEMMEN, M. MOHIYUDDIN, AND K. YELICK, *Avoiding communication in computing Krylov subspaces*, Tech. Report UCB/EECS-2007-123, EECS Dept., U.C. Berkeley, Oct 2007.

[8] D. GANNON AND J. VAN ROSENDALE, *On the impact of communication complexity on the design of parallel numerical algorithms*, Trans. Comput., 100 (1984), pp. 1180–1194.

[9] G. GOLUB AND C. VAN LOAN, *Matrix computations*, vol. 3, Johns Hopkins University Press, 2012.

[10] A. GREENBAUM, *Estimating the attainable accuracy of recursively computed residual methods*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 535–551.

[11] A. GREENBAUM AND Z. STRAKOŠ, *Predicting the behavior of finite precision Lanczos and conjugate gradient computations*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 121–137.

[12] A. Hindmarsh and H. Walker, *Note on a Householder implementation of the GMRES method*, Tech. Report UCID-20899, Lawrence Livermore National Lab., CA., 1986.
[13] M. Hoemmen, *Communication-avoiding Krylov subspace methods*, PhD thesis, EECS Dept., U.C. Berkeley, 2010.
[14] W. Joubert and G. Carey, *Parallelizable restarted iterative methods for nonsymmetric linear systems. Part I: theory*, Int. J. Comput. Math., 44 (1992), pp. 243–267.
[15] C. Leiserson, S. Rao, and S. Toledo, *Efficient out-of-core algorithms for linear relaxation using blocking covers*, J. Comput. Syst. Sci. Int., 54 (1997), pp. 332–344.
[16] G. Meurant, *The Lanczos and conjugate gradient algorithms: from theory to finite precision computations*, SIAM, 2006.
[17] G. Meurant and Z. Strakoš, *The Lanczos and conjugate gradient algorithms in finite precision arithmetic*, Acta Numer., 15 (2006), pp. 471–542.
[18] M. Mohiyuddin, *Tuning Hardware and Software for Multiprocessors*, PhD thesis, EECS Dept., U.C. Berkeley, May 2012.
[19] M. Mohiyuddin, M. Hoemmen, J. Demmel, and K. Yelick, *Minimizing communication in sparse matrix solvers*, in Proc. ACM/IEEE Conference on Supercomputing, 2009.
[20] C. Paige, M. Rozioznik, and Z. Strakoš, *Modified Gram-Schmidt (MGS), least squares, and backward stability of MGS-GMRES*, SIAM J. Matrix Anal. Appl., 28 (2006), pp. 264–284.
[21] Chris C Paige, *Accuracy and effectiveness of the lanczos algorithm for the symmetric eigenproblem*, Linear algebra and its applications, 34 (1980), pp. 235–258.
[22] G. Sleijpen and H. van der Vorst, *Reliable updated residuals in hybrid Bi-CG methods*, Computing, 56 (1996), pp. 141–163.
[23] E. Sturler, *A performance model for Krylov subspace methods on mesh-based parallel computers*, Parallel Comput., 22 (1996), pp. 57–74.
[24] S. Toledo, *Quantitative performance modeling of scientific computations and creating locality in numerical algorithms*, PhD thesis, MIT, 1995.
[25] C. Tong and Q. Ye, *Analysis of the finite precision bi-conjugate gradient algorithm for nonsymmetric linear systems*, Math. Comp., 69 (2000), pp. 1559–1576.
[26] H. Van der Vorst and Q. Ye, *Residual replacement strategies for Krylov subspace iterative methods for the convergence of true residuals*, SIAM J. Sci. Comput., 22 (1999), pp. 835–852.
[27] J. Van Rosendale, *Minimizing inner product data dependencies in conjugate gradient iteration*, Tech. Report 172178, ICASE-NASA, 1983.
[28] H. Walker, *Implementation of the GMRES method using Householder transformations*, SIAM J. Sci. Stat. Comput., 9 (1988), pp. 152–163.
[29] J-P. Zemke, *Krylov subspace methods in finite precision: a unified approach*, PhD thesis, Dissertation, Technische Universität Hamburg-Harburg, Hamburg, Germany, 2003.